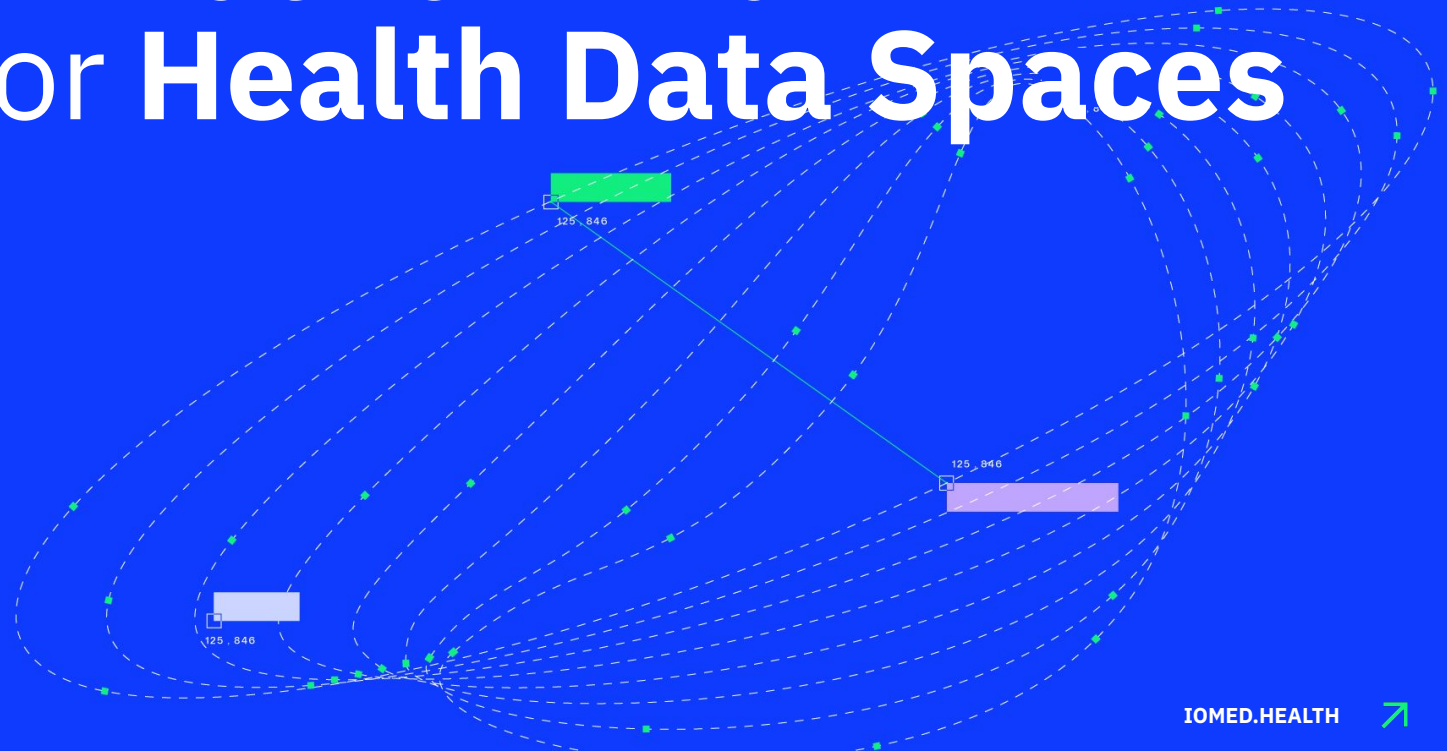
IOMED

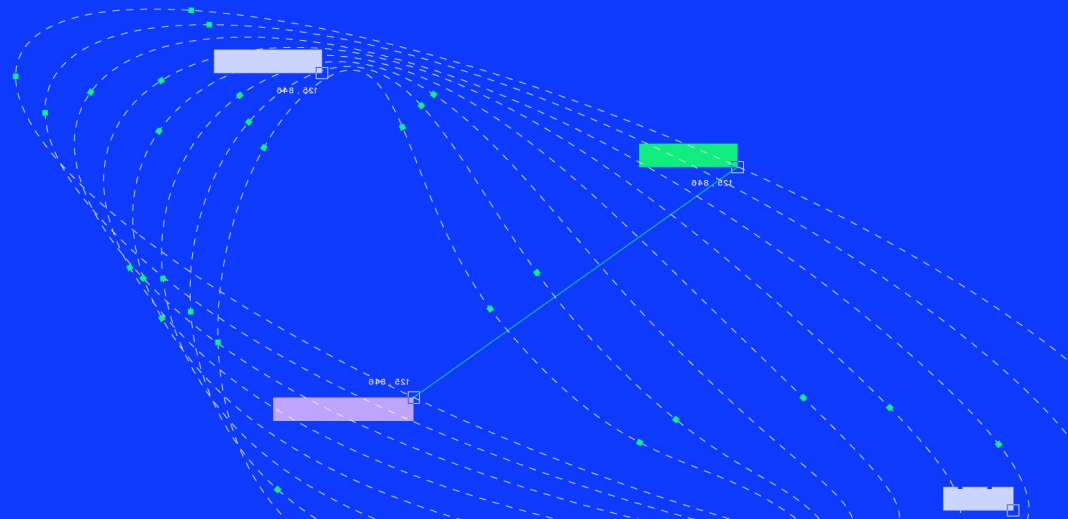
#3626

A comprehensive Evaluation Framework for **Artificial Intelligence** in Clinical Data Extraction And Normalization For **Health Data Spaces**



Gabriel Maeztu MD *et al.*
CTO - IOMED R&D Department
ISPOR 2026 - Philadelphia

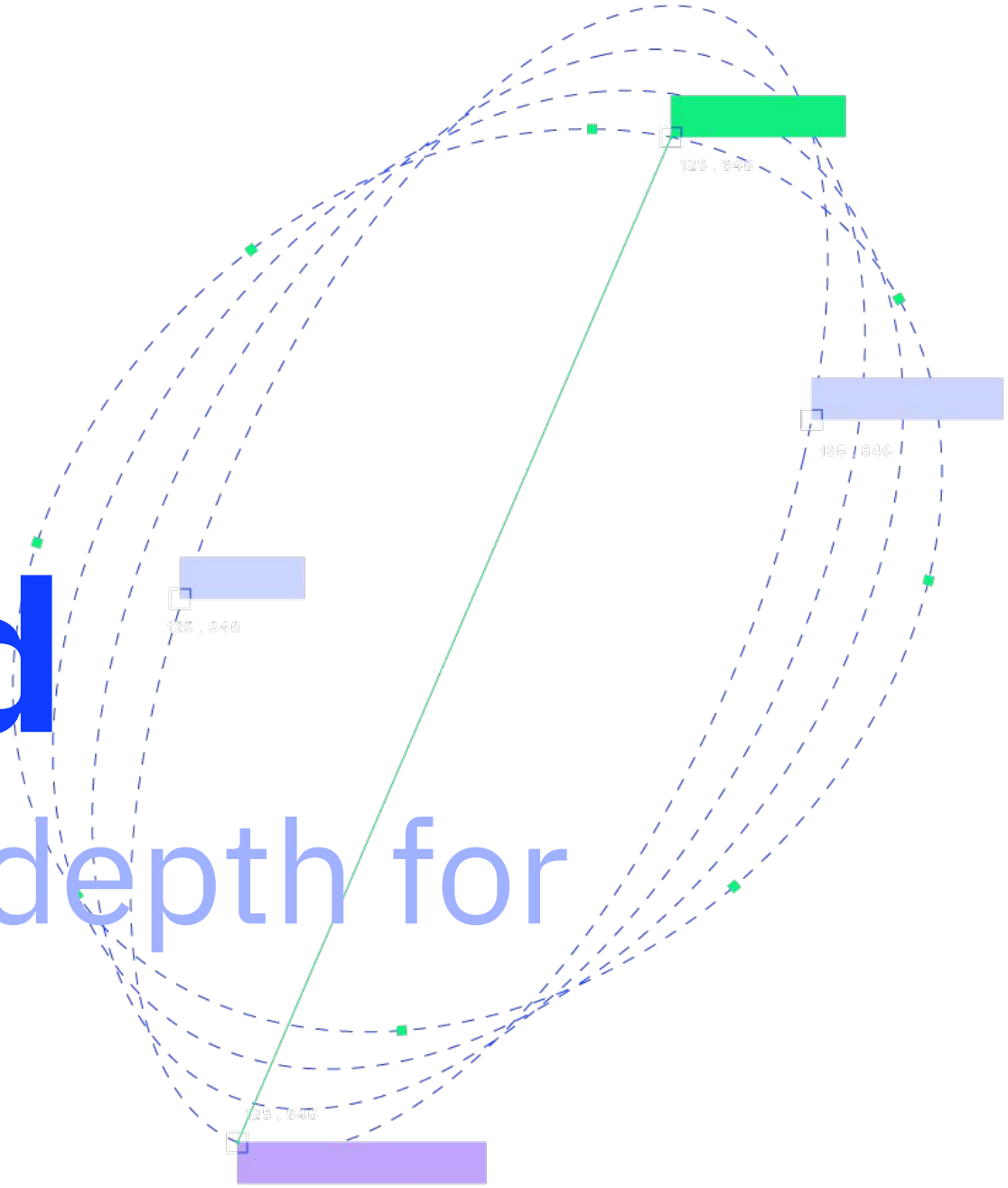
Index



[01]

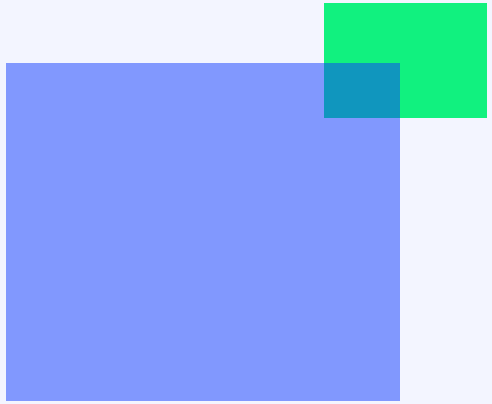
Background

The challenge of depth for RWE generation



IOMED

Healthcare and Lifesciences have long had an **uneven data Demand/Supply** relationship[1].



SUPPLY

Providers are only able to supply small, **shallow or inconsistently** collected dataset.

DEMAND

Need for **datasets with clinical depth**, collected in the least possible time.

Approximately **60% to 80% of real world data (RWD) exists in unstructured formats,** complicating direct analysis and scalability for Real-World Evidence (RWE) studies[2]

Traditional **manual mapping of hospital data to standard concepts is labor-intensive**, error-prone, and hinders the secondary use of data.

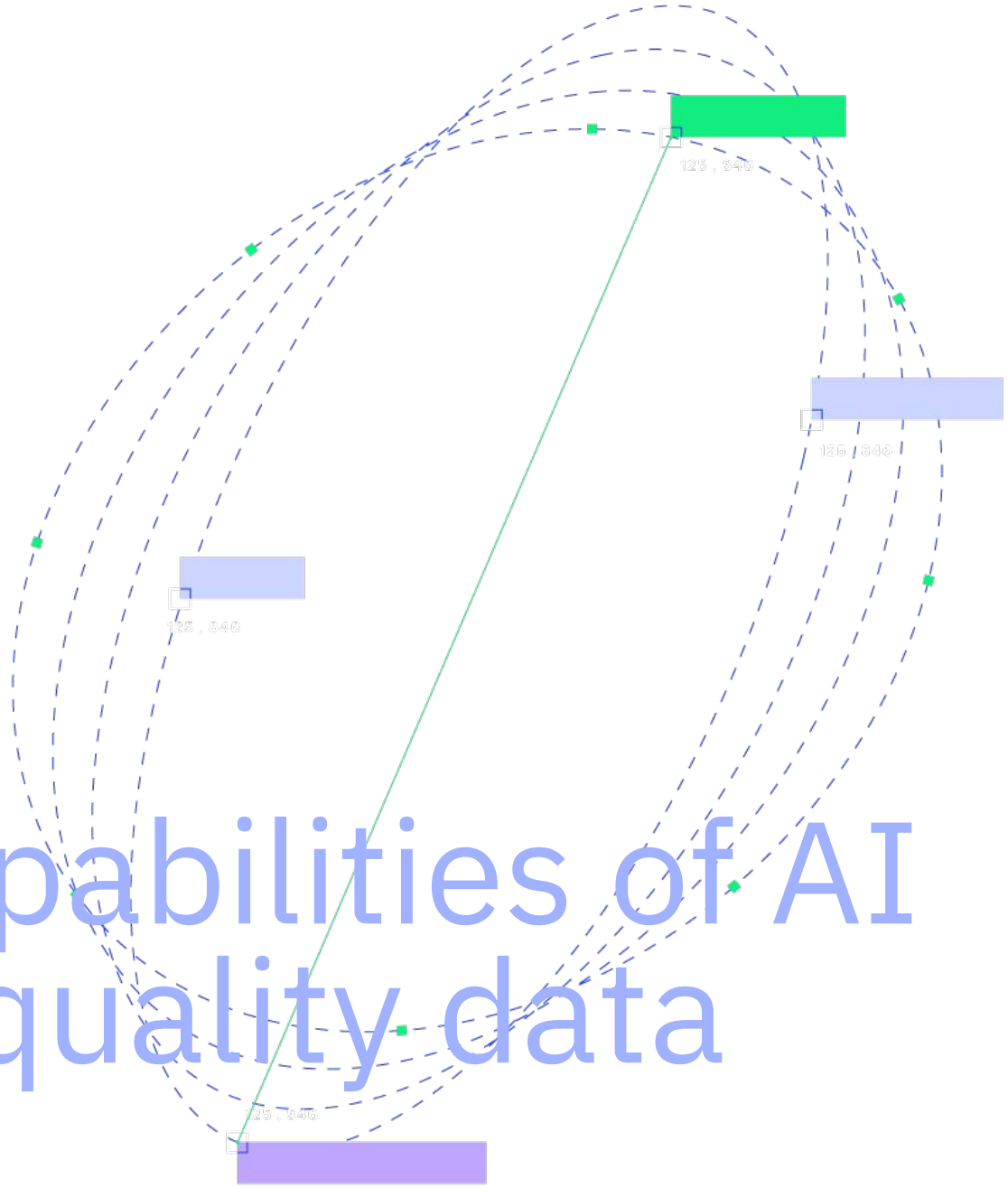
Leveraging **Artificial Intelligence (AI)**, **to bridge the gap between** raw narratives and standardized health data.

[1]: Grimberg F, Asprion PM, Schneider B, Miho E, Babrak L, Habbabeh A. The Real-World Data Challenges Radar: A Review on the Challenges and Risks regarding the Use of Real-World Data. Digit Biomark. 2021 Jun 24;5(2):148-157. doi: 10.1159/000516178. PMID: 34414352; PMCID: PMC8339486.
[2]: Zhao X et al. Integrating real-world data to accelerate and guide drug development: A clinical pharmacology perspective. Clin Transl Sci. 2022 Oct;15 (10):2293–2302.

[02]

Objectives

Evaluating the capabilities of AI to generate high quality data



IOMED

Objective and Scope of Evaluation

Objective:

Establish a comprehensive **verification and validation (V&V) framework for** research grade data structured with **AI**. Specifically addressing 2 critical tasks to capture/normalize of data:

- Automatic Term Mapping (ATM)

  ⇒ Normalizing diverse codifications into standard (*e.g. Internal code to ICD 10*)

- Natural Language Processing (NLP)

  ⇒ Extracting relevant data from clinical narratives (*e.g. diagnoses, procedures..)*

Secondary Aims:

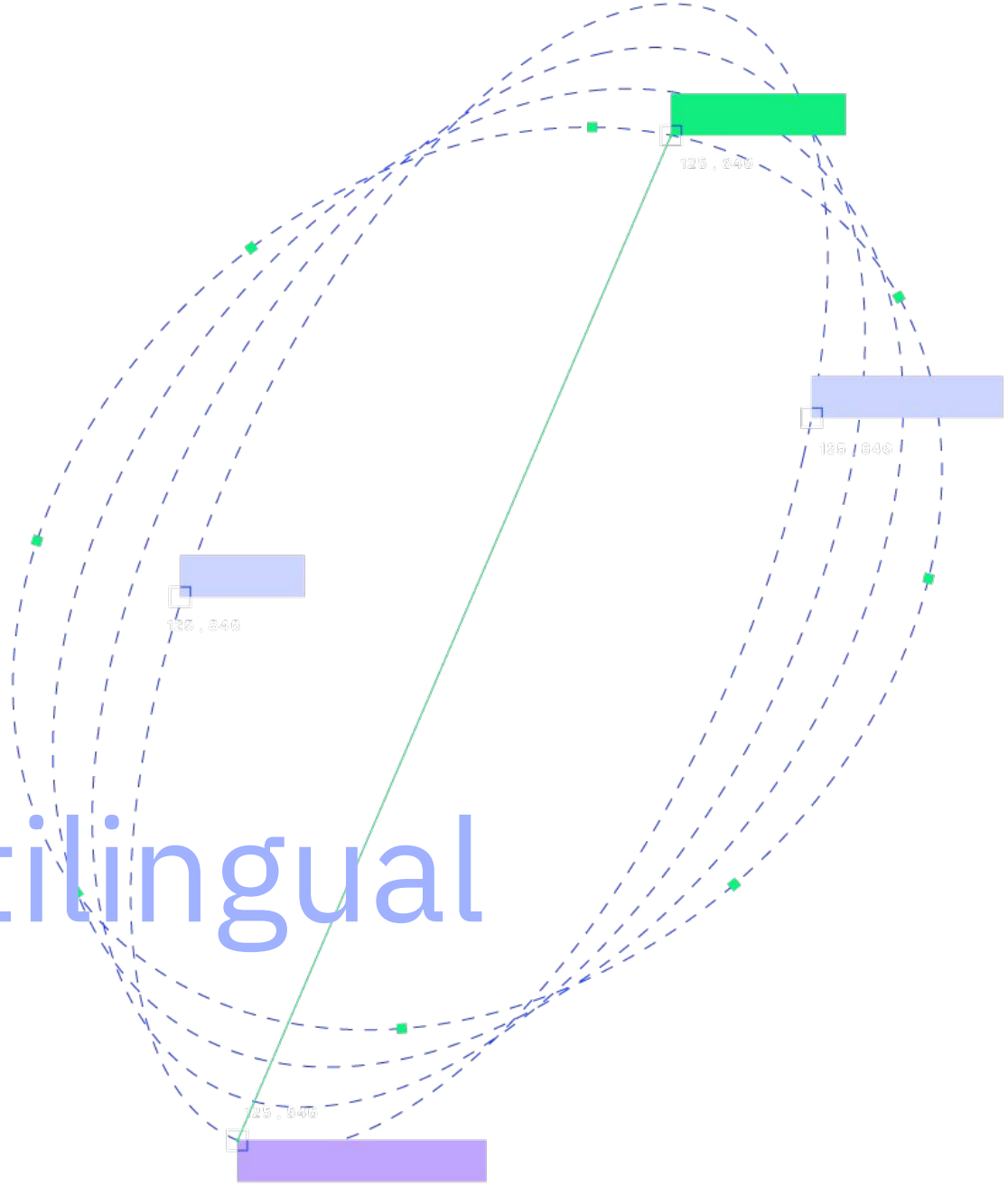
Demonstrate the **generalizability** and **reproducibility** of the AI capabilities across:

- x11 Therapeutic areas
- x10 Geographic regions (*within Europe*)
- x3 Languages

[03]

Materials

Multicentric, multilingual RWD studies



IOMED

RWD Datasets with and without AI

10 multicentric RWD database studies were conducted and analyzed. For each study, a comparative analysis was performed between datasets **with and without the integration of AI**-generated data points:

- **Arm A**: Datasets utilizing only pre-existing structured data and manual mappings *(e.g., ICD-10, SNOMED CT, LOINC)*
- **Arm B:** Datasets enhanced with AI-generated data points *(e.g. Internal coding, Clinical notes)*

Each study utilized 3 to 7 of **10 unique clinical databases** from European tertiary Hospitals.

550 unique variables were evaluated across the different studies.

The evaluation encompassed **diverse therapeutic areas** across diverse clinical conditions, including:

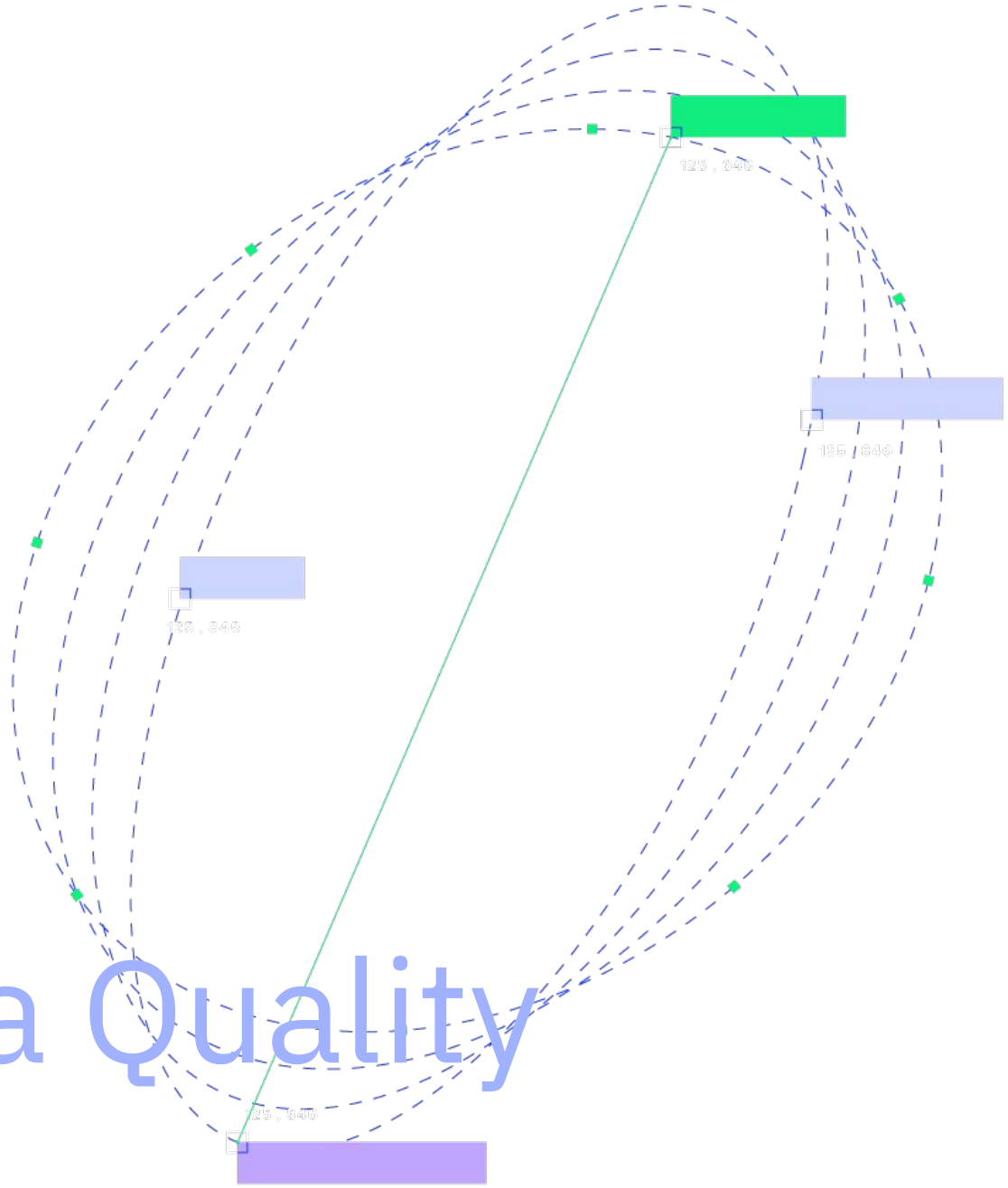
- **Oncology:** Multiple Myeloma, Acute Myeloid Leukemia (AML), Prostate Cancer, and Endometrial Cancer.

- **Chronic & Rare Diseases:** Multiple Sclerosis, Chronic Urticaria, Psoriasis, Ulcerative Colitis, Hypophosphatemia, Alport Syndrome and Chronic Kidney Disease

[04]

Methods

Micro and macro evaluation of Data Quality



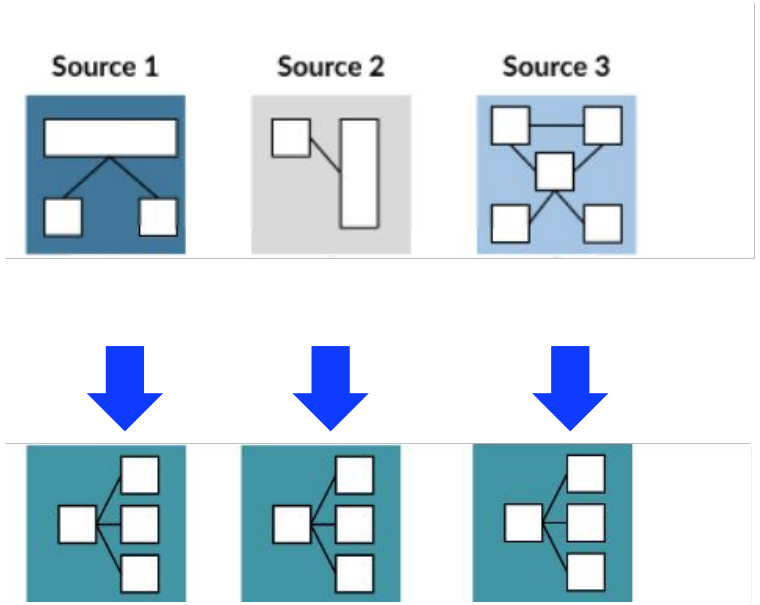
IOMED

A reproducible approach thanks to Data Standardization

All the **databases were mapped to the OHDSI OMOP Common Data Model (CDM)**[1] 5.4 for interoperability, mapping between 5 and 8 information systems at each Hospital (e.g. EHR, LIMS, PMS, Microbiology..) into the standard

All the studies were executed using the EMA's DARWIN EU[2] tooling for analytics.

Once the databases where ready for analysis, AI was applied to extend the content. AI was iteratively improved in multiple rounds thanks to annotated examples (few-shots learning).



OMOP CDM Format

[1]: Voss EA, Makadia R, Matcho A, Ma Q, Knoll C, Schuemie M, DeFalco FJ, Londhe A,Zhu V, Ryan PB. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. J Am Med Inform Assoc. 2015 May;22(3):553-64. doi: 10.1093/jamia/ocu023. Epub 2015 Feb 10. PubMed PMID: 25670757; PubMed Central PMCID: PMC4457111
[2]: Dernie F, Corby G, Robinson A, Bezer J, Mercade-Besora N, Griffier R, Verdy G, Leis A, Ramirez-Anguita JM, Mayer MA, Brash JT, Seager S, Parry R, Jodicke A, Duarte-Salles T, Rijnbeek PR, Verhamme K, Pacurariu A, Morales D, Pinheiro L, Prieto-Alhambra D, Prats-Uribe A. Standardised and Reproducible Phenotyping Using Distributed Analytics and Tools in the Data Analysis and Real World Interrogation Network (DARWIN EU). Pharmacoepidemiol Drug Saf. 2024 Nov;33(11):e70042. doi: 10.1002/pds.70042. PMID: 39532529.

IOMED

Structuring data with AI, Reading Clinical Narratives

Multiple AI tasks were performed in order to extract more information in each of the databases.

NLP was used to handle the "*unstructured*" nature of medical records (notes, reports) by transforming them into a standardized, research-ready codes in OMOP CDM.

Tasks:

- **Identifying Entities:** Finding diseases, medications, and procedures within free text.
- **Capturing Context:** Determining if the event are current, past, negated (e.g., "patient denies chest pain"), refer to patient...

noteID	person_id	date	value	visit
1	243	2021-10-12	follow up, 02-17-2015 usual f/up Follow-Up: Consultation on pian left arm since yesterday. Thi smorning pain persists. Had similar pain whn had AMI. Has taken nitrostat without improvement. Request of ECG	367

Data stored in free-text format

...	person_id	procedure_concept_id	procedure_date	visit_occurrence_id	...
	243	4115169	2021-10-12	896	
	..	...	...	...	...

Pain in the left arm

OMOP CDM
Procedure Occurrence

IOMED

Structuring data with AI, Mapping codes to standard terminologies

ID	patient_id	data_id	date	value	unit	visit
643	243	weight	2021-10-12	190	lbs	367
2891	243	weight	2021-11-21	170	lbs	458

Non-standard codification

Automapping model

...	person_id	observation_concept_id	observation_date	value_as_number	...
...	243	3025315	2021-10-12	86.18	...

OMOP CDM Observation

ATM was used to handle the "*internal coding*" of events in the databases, transforming them into a standardized, research-ready codes in OMOP CDM.

This AI capability, given a internal representation of a datapoint is able to find the corresponding code *(e.g. ICD10, LOINC)* in a standard terminology or ontology.

IOMED

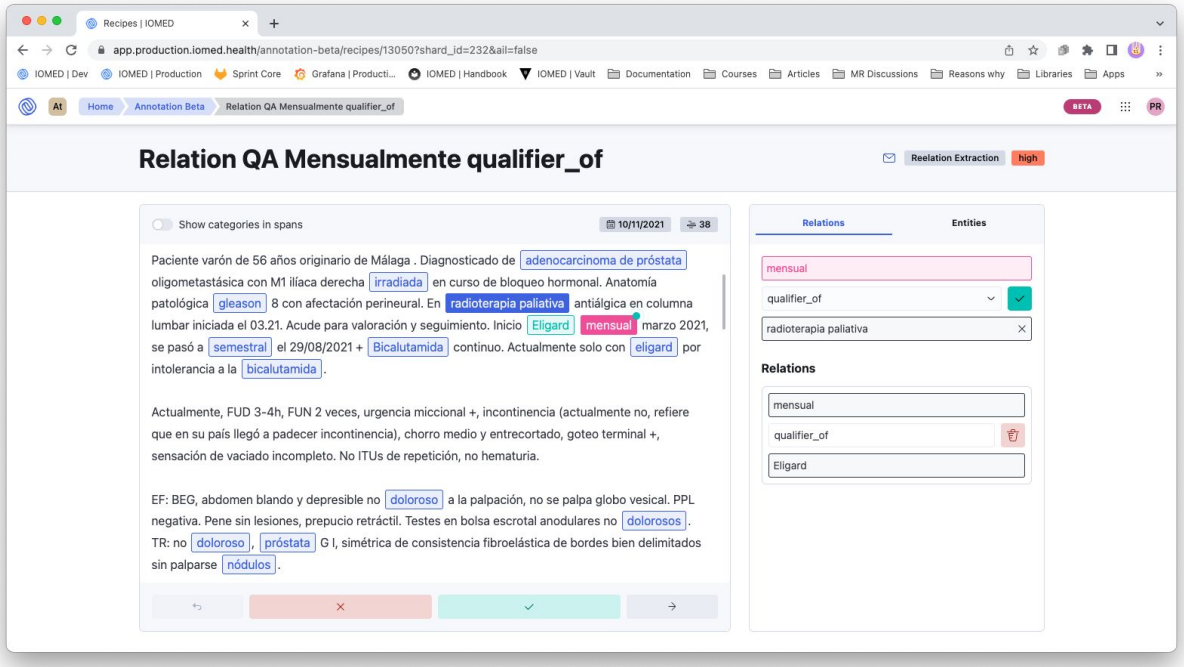
Ensuring Trust: The Physician-Led Verification Framework

A **remote Source Data Verification (rSDV)** process where physicians manually double-check **AI-extracted data against the original clinical** source. This process yield metrics to evaluate the TP, FP, TN, but not FN.

Using Bayesian statistics to determine the exact number of records a physician needs to review for each variable to achieve a 95% confidence level.

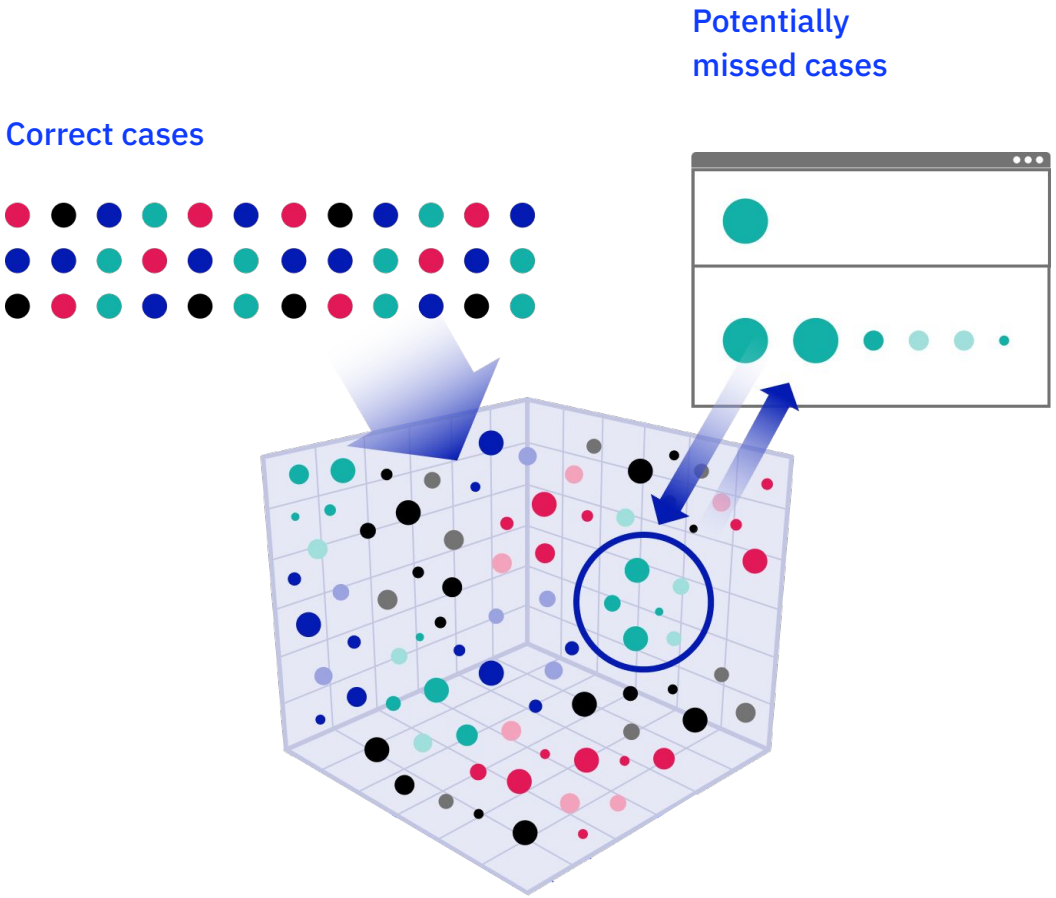
Inter-rater Reliability Analysis

We conducted a comprehensive inter-rater reliability analysis. This analysis is crucial for evaluating the performance of physician performing the annotation in the verification process, which serves as a critical quality control measure for our AI capabilities.



IOMED

Ensuring Trust: The Physician-Led Verification Framework



Identifying **False Negatives** (missed relevant clinical information) is difficult due to their rarity in large, unannotated datasets.

Random sampling is ineffective, and manual review of all data is impractical.

We address this using a **semantic similarity search approach**[1], generating query vectors from correctly identified positive cases. These vectors capture semantic characteristics and are compared against the entire dataset to find similar, previously unrecognized instances.

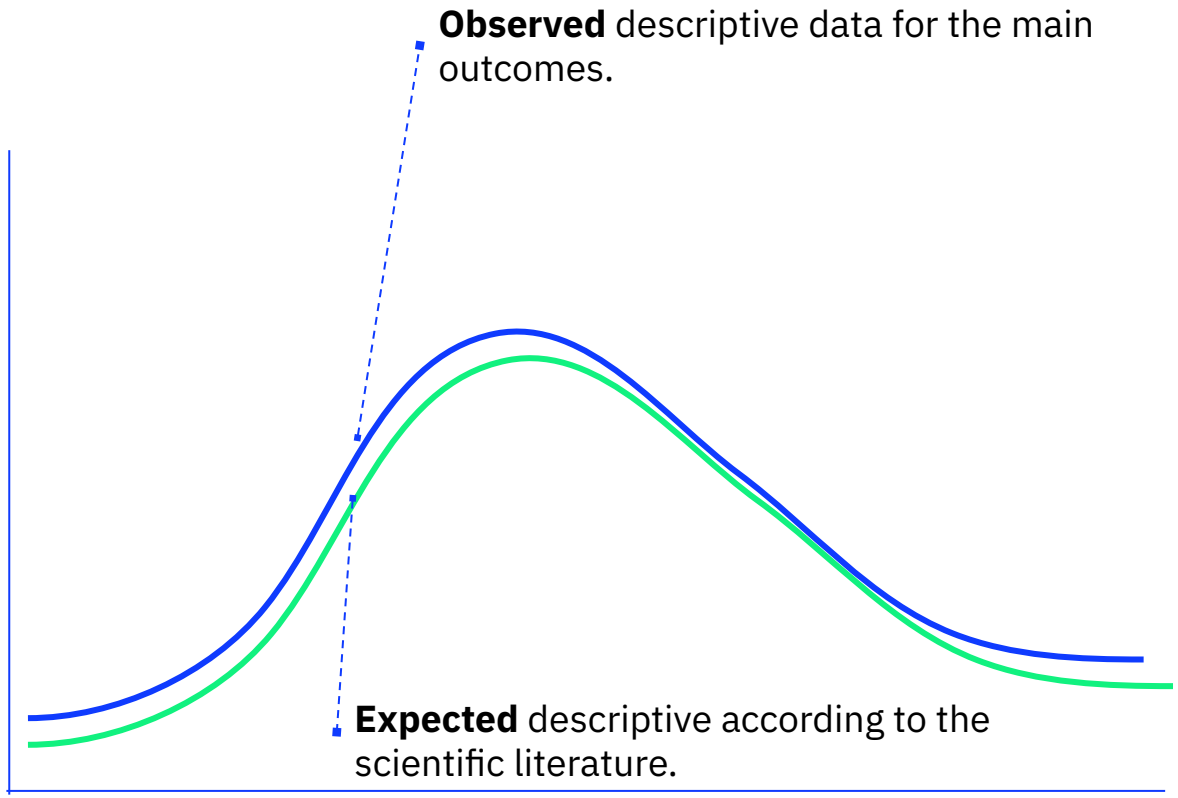
[1]: Maria Quijada, Maria Vivó, Álvaro Abella-Bascarán, Paula Chocrón, and Gabriel de Maeztu. 2022. A Framework for False Negative Detection in NER/NEL. In Natural Language Processing and Information Systems: 27th International Conference on Applications of Natural Language to Information Systems, NLDB 2022, Valencia, Spain, June 15–17, 2022, Proceedings. Springer-Verlag, Berlin, Heidelberg, 323–330. https://doi.org/10.1007/978-3-031-08473-7_30

IOMED

Ensuring Trust: The Data Scientist-Led Validation

A systematic audit of the entire dataset based on the Kahn Framework[1] to ensure three key qualities:

- **Completeness:** Are there missing pieces of the patient story?
- **Conformance:** Does the data follow the required medical standards
- **Plausibility:** Does the data make medical sense (e.g., ensuring no diagnosis dates precede birth dates)?

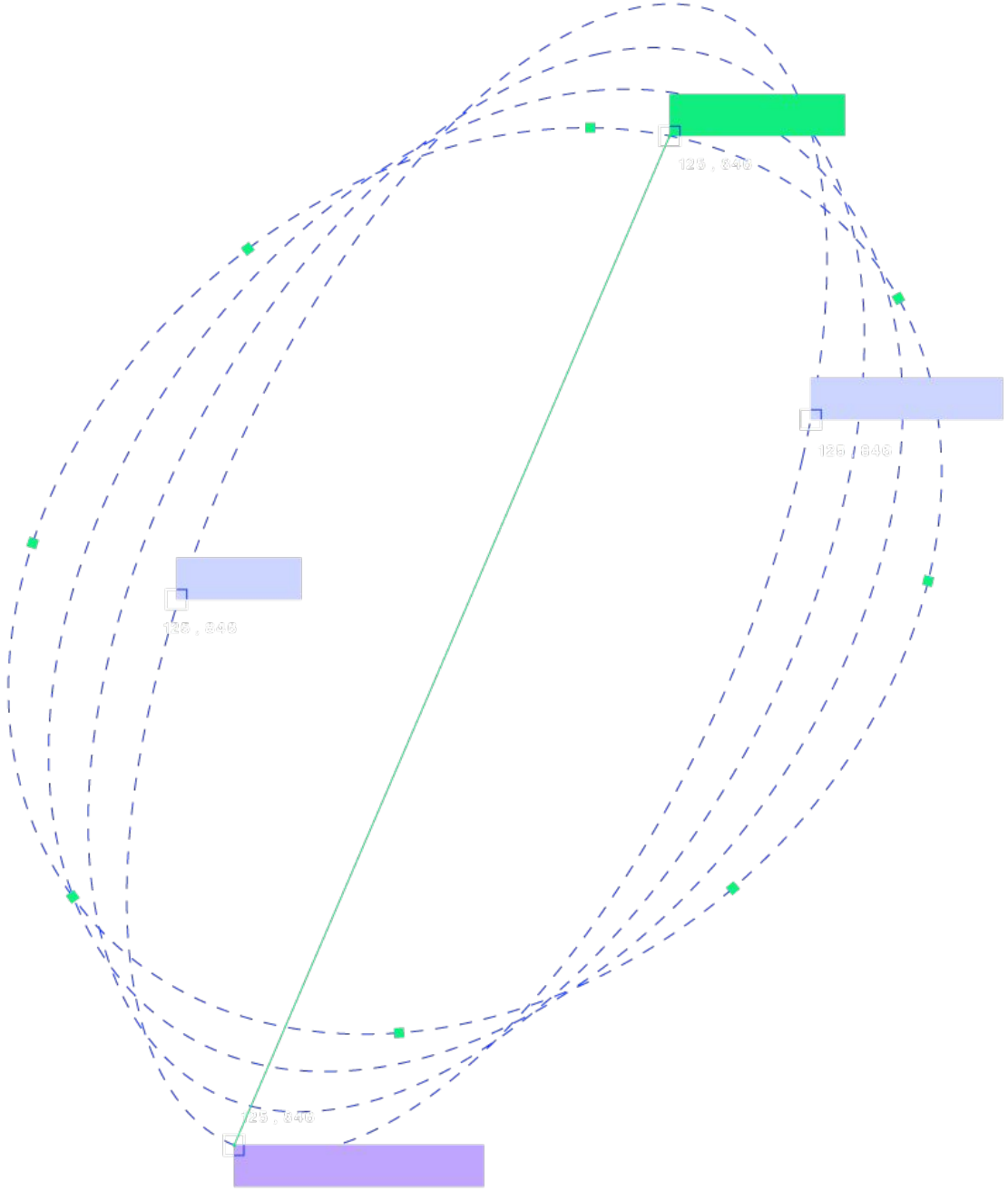


[1]:Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, Estiri H, Goerg C, Holve E, Johnson SG, Liaw ST, Hamilton-Lopez M, Meeker D, Ong TC, Ryan P, Shang N, Weiskopf NG, Weng C, Zozus MN, Schilling L. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. EGEMS (Wash DC). 2016 Sep 11;4(1):1244. doi: 10.13063/2327-9214.1244. PMID: 27713905; PMCID: PMC5051581.

[05]

Results

The challenge



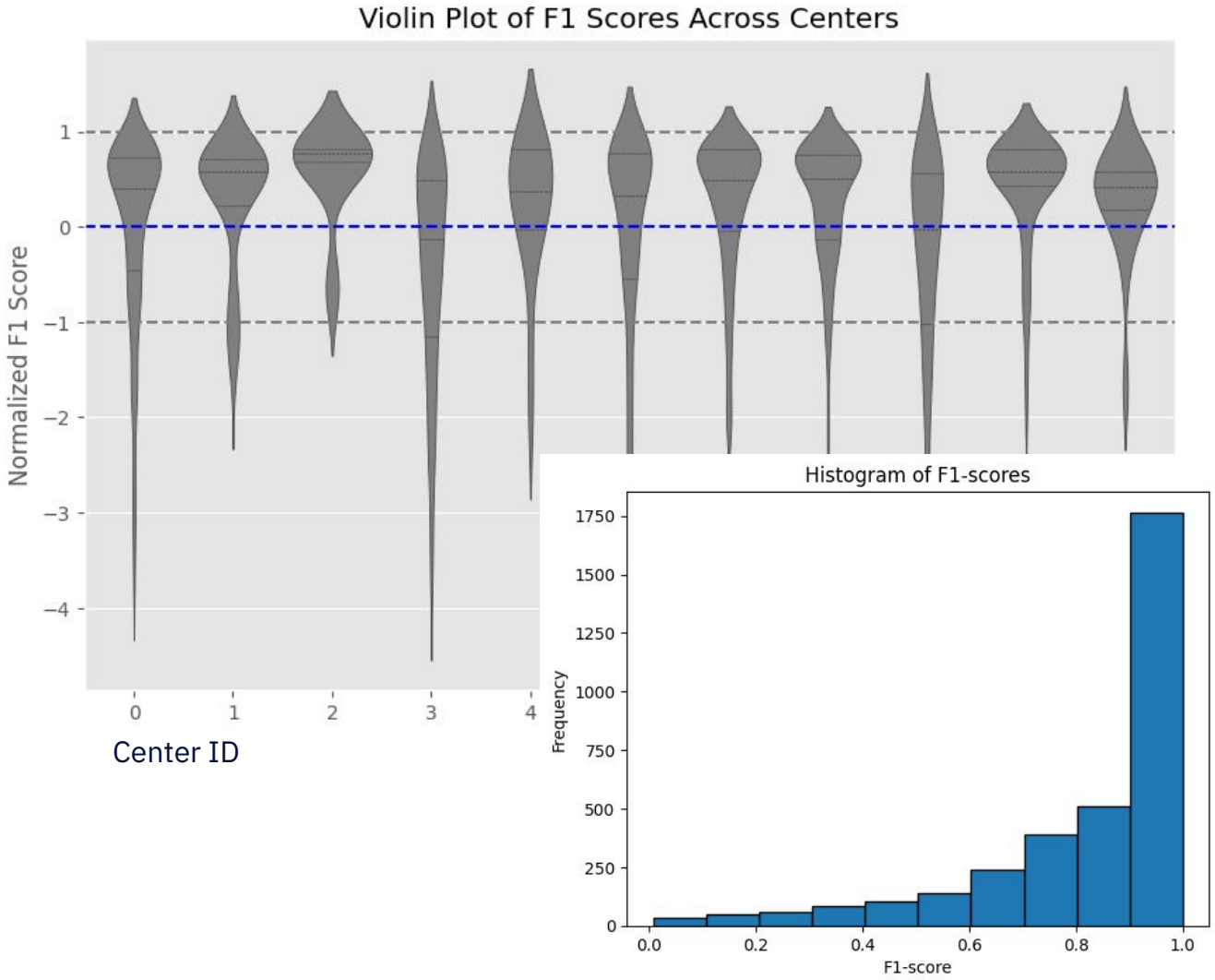
IOMED

AI Performance Accuracy and Center Consistency

Inference Accuracy for NLP:
Achieved a pooled **F1-score of 0.88** (95% CI: 0.84-0.92) across all 10 centers and 10 studies.

Center Generalizability:
Intraclass Correlation Coefficient (ICC) of 0.0014 for F1-scores, indicating minimal between-center variability and high consistency across diverse settings.

Heterogeneity:
Levene's Test confirmed that while variances differed, performance remained robust across different hospitals.



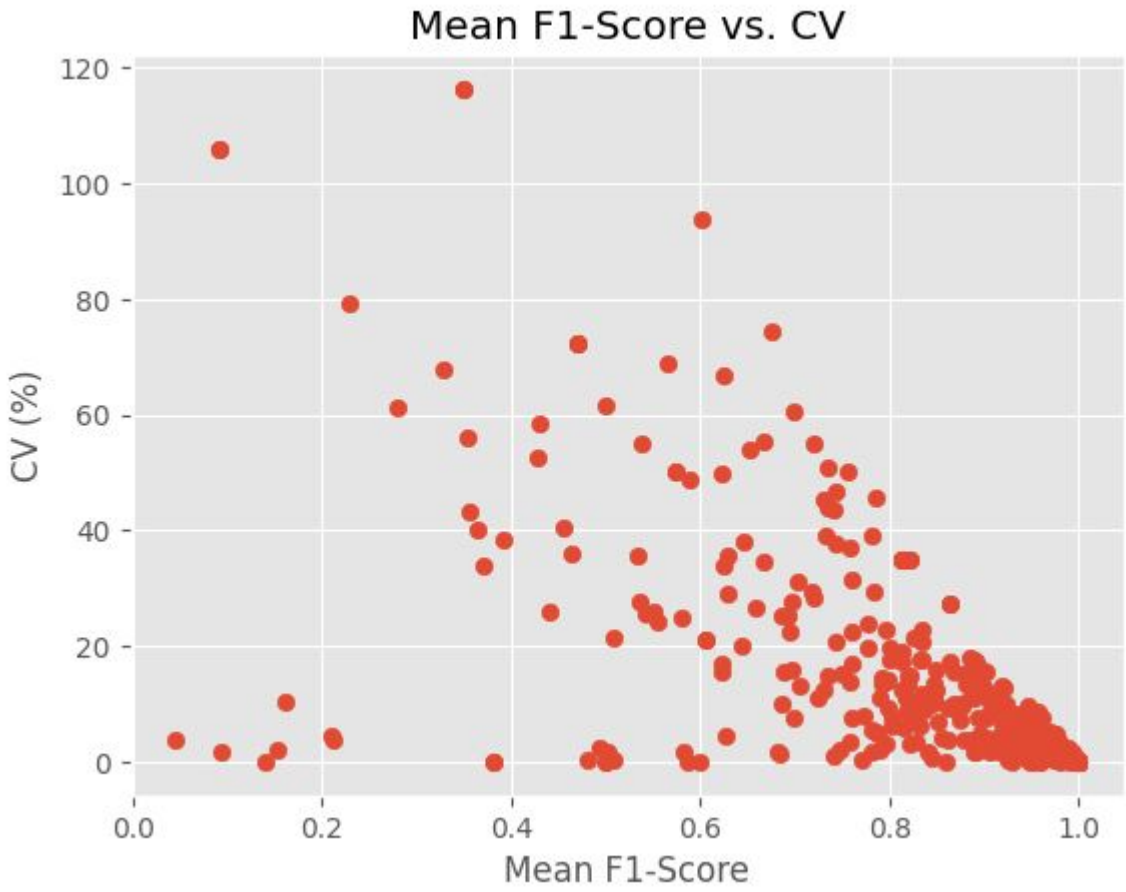
IOMED

AI Performance Accuracy and Center Consistency

The Coefficient of Variation (CV) was used to assess the relative variability of F1 scores yielding a value of 0.2557 (25.57%).

This indicates a moderate, but manageable, level of variability relative to the mean F1 score.

Furthermore, as illustrated in the scatter plot, **as the AI models were further trained and achieved higher Mean F1-Scores**, the variation was significantly reduced, demonstrating greater stability.



IOMED

Impact on Data Volume and Diversity

Volume Increase:
Integration of AI-generated data points resulted in a 40% increase in unique data points (from 1.2M to 1.68M)

Diversity Enrichment:
Margalef's Index of diversity significantly improved from an average of 3.5 to 5.8 across centers.

Record Density:
The median number of recorded concepts per patient per domain increased from 5 to 9 ($p < 0.001$).

Temporal Coverage:
Patient records saw an average 12-month extension in temporal span.

Granularity:
Event density per patient per year increased from 3.2 to 6.5 events, providing a more detailed longitudinal view.

Accuracy Benchmark:
Physicians demonstrated a 91% agreement rate when compared against golden annotators.

IOMED

Downstream Clinical Outcome Analysis

Impact on the final outcomes
Inclusion of AI-generated data significantly impacted primary clinical outcomes across cohorts ($p < 0.05$). Cohen's d effect sizes ranged from 0.75 to 0.95, indicating a substantial impact on study results

Improved Sensitivity:
AI-enhanced datasets showed improved odds ratios, suggesting a higher likelihood of capturing positive clinical events.2

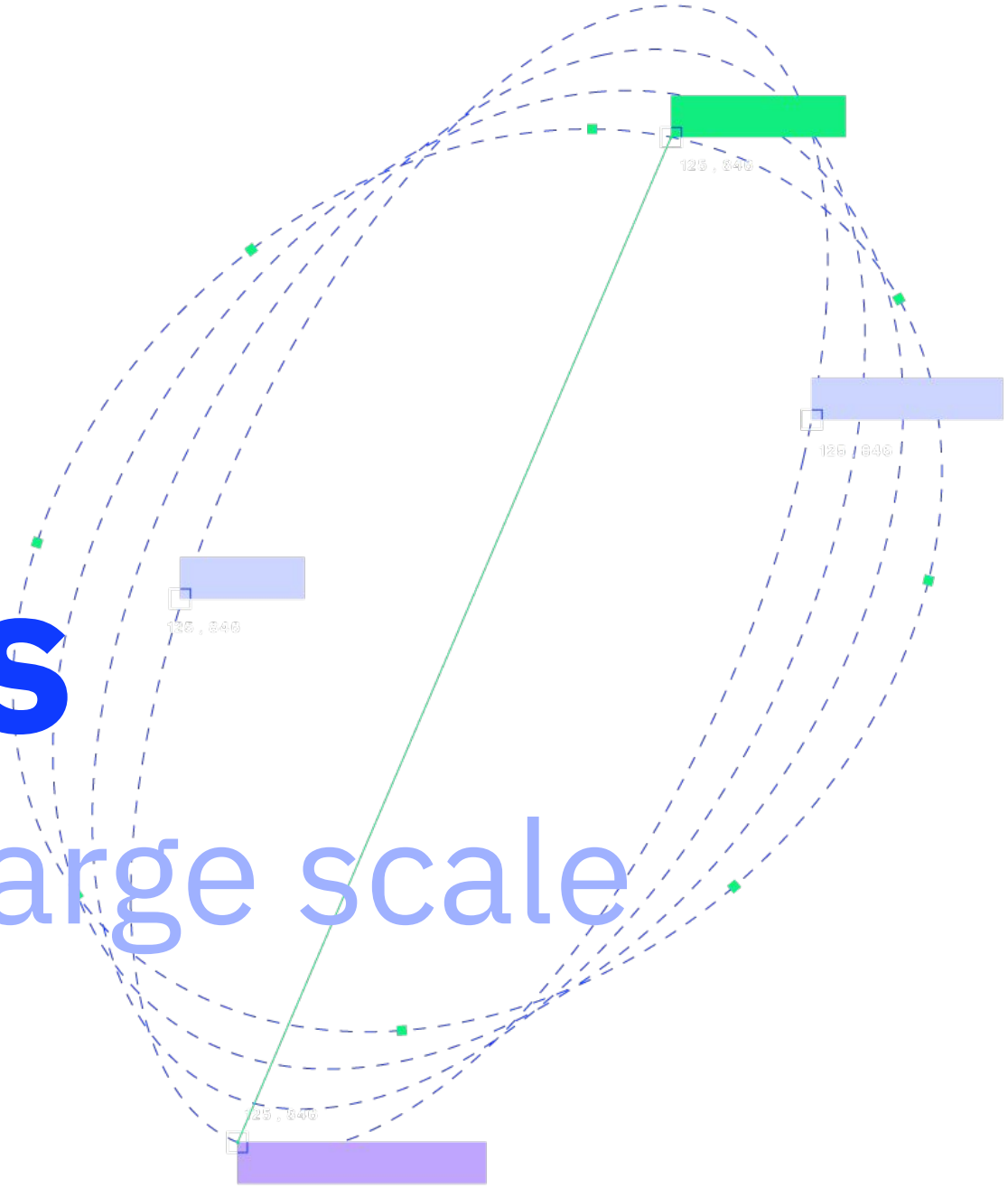
Table 3. Effect Size and Odds Ratios per Cohort

Cohort	Cohen's *d*	Odds Ratio
Multiple Sclerosis	0.85	1.2
Chronic Urticaria	0.90	1.5
Psoriasis	0.78	1.8
Ulcerative Colitis	0.95	2.1
Multiple Myeloma	0.82	1.9
Acute Myeloid Leukemia	0.75	1.7
Prostate Cancer	0.88	2.0
Endometrial Cancer	0.92	2.2
Hypophosphatemia	0.80	1.6
Alport Syndrome	0.84	2.1

[06]

Conclusions

AI readiness for large scale RWE



IOMED

Validation of the Framework and Outcomes

- The implementation of a rigorous V&V framework confirms that **AI-generated clinical data can be reliable for RWE**.
- The framework proved **robust generalisation** across 10 centers and 10 therapeutic areas, demonstrating its readiness for scale.
- We **improved the predictive power and likelihood of clinical outcomes.**

IOMED

Lessons learned

- AI must be adapted to each Health Data Space. High CV's are the norm for the variables when the algorithms are not further trained. Iterative training reduced inter-centre and inter-cohort CV.

- Inter-annotator evaluation is necessary to avoid biased AI evaluation.

IOMED

Gabriel Maeztu MD
gabi@iomed.health

C/ de St. Antoni Maria Claret, 167, Sant Leopold Pavillion, 08025 — Barcelona

info@iomed.health

IOMED.HEALTH